Predicting perceptions of the lexical richness of short French, German, and Portuguese texts using text-based indices

Jan Vanhove, Audrey Bonvin, Amelia Lambelet & Raphael Berthele

Institute of Multilingualism, Fribourg | Switzerland

Abstract: We investigated how well readers' perceptions of the lexical richness of short texts can be predicted on the basis of automatically computable indices of the texts' lexical properties. 3,060 French, German and Portuguese texts (between 9 and 284 words long) written by 8- to 10-year-olds were rated for their lexical richness by between 3 and 18 uninstructed raters, and over 150 indices were derived from these texts. We found that the ratings could to a substantial degree be predicted on the basis of these indices and that the accuracy with which the ratings of shorter texts could be predicted was comparable to that of longer texts. For French and German, the greatest predictive power was attained by opaque models with scores of predictors, but models with fewer predictors based on a 6-dimensional framework of lexical richness perception or even with a single, easily computed predictor, Guiraud's index, fared only slightly worse.

Keywords: human ratings, lexical diversity, lexical sophistication, lexical richness, predictive modelling

journal of WRITING RESEARCH

Vanhove, J., Bonvin, A., Lambelet, A., & Berthele, R. (2019). Predicting perceptions of the lexical richness of short French, German, and Portuguese texts using text-based indices. Journal of Writing Research, 10(*3*), 499-525. https://doi.org/10.17239/jowr-2019.10.03.04

Contact: Jan Vanhove, University of Fribourg, Department of Multilingualism, Rue de Rome 1, Fribourg, 1700 | Switzerland – jan.vanhove@unifr.ch.

1. Introduction

The present study is about how and how well perceptions of written texts can be predicted on the basis of quantifiable characteristics of those texts. Its backdrop is a substantial body of research that has investigated the textual characteristics that affect expert ratings of text quality. In addition to the number of formal errors in the texts and the use of syntactic and cohesive devices, these characteristics include text length, the complexity and sophistication of the words used in the text, and the variety of the words in the text. Of particular relevance to the present study, metrics related to the quantity, range, and variety of the vocabulary used in a text—that is, metrics related to the text's *lexical richness*—are important predictors of writing quality as judged by experts (e.g., Crossley & McNamara, 2011; Engber, 1995; Grobe, 1981; Jarvis, 2002; Jarvis, Grant, Bikowski, & Ferris, 2003; Malvern, Richards, Chipere, & Durán, 2004; McNamara, Crossley, & McCarthy, 2010; Nold & Freedman, 1977).[1]

Yet, as Jarvis (2013a) points out, despite the proliferation of lexical richness metrics in quantitative linguistics, it is not clear how well these do indeed measure aspects of lexical richness. To elaborate, lexical richness metrics are generally evaluated on the basis of how sensitive they are to text length (e.g., McCarthy & Jarvis, 2007, 2013; Treffers-Daller, 2013; Tweedie & Baayen, 1998) and how well they correlate with variables representing other constructs, such as essay quality, language proficiency or other writer and talker characteristics (e.g., Daller, van Hout, & Treffers-Daller, 2003; Treffers-Daller, 2013; Treffers-Daller, Parslow, & Williams, 2016). But what is unclear is how well any given lexical richness metric measures what it is intended to measure: some aspect of lexical richness. As Jarvis (2013a) writes,

> "the problem is not that the existing measures fail to predict language proficiency, aphasia, and so forth; the problem is that they lack construct validity because they have not been derived from a well-developed theoretical model of lexical diversity [i.e., what we call lexical richness]." (p. 95)

Jarvis (2013a) suggests that progress can be made by validating lexical richness metrics not in terms of how well they avoid text-length artefacts or how well they are correlated to manifestations of other constructs, but "in accordance with their ability to predict the lexical diversity [i.e., lexical richness] judgments of human raters" (p. 101). That is, he suggests that lexical richness be treated as a perceptual phenomenon, and more specifically, he argues that it is the judgement of *untrained* human raters that should be the touchstone of lexical richness metrics. In addition, Jarvis (2013a, 2013b, 2017) proposes that the perceptual phenomenon of lexical richness may be captured in a limited number of dimensions, which we outline in the next section.

The present study contributes to the research programme proposed by Jarvis. We collected data on the lexical richness of 3,060 French, German, and Portuguese texts written by third and fourth graders as perceived by untrained native speakers of the

respective languages. This database supplements a literature on human ratings dominated by English texts written by older students and judged by experts. For each of these texts, we computed a large number of indices that were conceivably related to these lexical richness perceptions. These indices were then used to build statistical models capable of predicting a text's perceived lexical richness. Our chief goals in doing so were (a) to assess how well perceptions of lexical richness can be predicted using text-based indices and (b) to gauge to what extent Jarvis' dimensional framework of perceived lexical richness trades off predictive power for theoretical interpretability. Our results suggest, first, that perceptions of lexical richness can to a substantial degree be predicted using text-based indices, even when the raters are not trained, and, second, that the trade-off between predictive power and theoretical interpretability entailed by Jarvis' dimensional framework is modest.

1.1 A dimensional framework of lexical richness

Jarvis (2013a, 2013b, 2017) suggests that the perceptual phenomenon of lexical richness may be captured in at least six dimensions: volume, variability, evenness, rarity, disparity, and dispersion.[2]

Volume refers to the texts' length, which can be expressed as the number of words they contain. Ordinarily, text length is seen as a confounding factor in research on lexical richness as most metrics of word repetition cannot sensibly be compared between texts of different length (e.g., Koizumi, 2012; Malvern et al., 2004). However, Jarvis' framework is concerned with how readers *perceive* lexical richness, and so is our study. The inclusion of the volume dimension reflects the possibility that readers may differentially perceive the lexical richness of otherwise comparable texts depending on their length.

Variability is the complement of word repetition. In other studies, this is often referred to as lexical diversity (see Note 1). A wide array of measures represent attempts to quantify variability, the most well-known of which is the type–token ratio (TTR), that is, the ratio of the number of unique words (types) to the number total words (tokens) in the text. Due to their design and the properties of language, most of these measures are, like the TTR, systematically related to text length such that they overlap with the first dimension, volume.

Evenness is intended to capture differences in the extent to which tokens of different types contribute to the text: Is there one type occurring regularly and other types all occurring rarely, or do different types all occur about equally frequently? Jarvis (2013b) uses the standard deviation of the counts of tokens per type as a first operationalisation of evenness.

Rarity concerns the frequency with which the words used in the text occur in the language at large. Previous research has expressed rarity as the proportion of words occurring in the text that do not belong to the *n* most frequent in the language (e.g., Laufer & Nation, 1995), the words' mean frequency in an external frequency list (e.g., Kyle & Crossley, 2015), or the words' mean frequency rank according to an external

frequency list (e.g., Jarvis, 2013b). In other studies, measures related to word frequency are often referred to as measures of lexical sophistication.

Disparity refers to how similar semantically or formally the words in the text are (Jarvis, 2013a, 2017). Jarvis (2013b) operationalises semantic similarity as the "mean number of words in the text that share the same semantic sense" according to WordNet. We are not aware of similar tools for French, German and Portuguese, so we will leave semantic disparity out of consideration. Jarvis (2013b) does not propose an operationalisation of formal disparity.

Dispersion, finally, pertains to whether tokens of the same type are distributed uniformly throughout the text or are clustered around the same place in the text. Jarvis (2013b) operationalised this as the average distance between different tokens of the same type, averaged over all types in the text, but currently, he computes it as the number of times that types are repeated within the next *n* (e.g., 20) tokens (personal correspondence, August 2, 2017). The total number of 'close repeats' is then divided by the total number of tokens. For this computation, the top-5 most frequent types in the language are not taken into account.

1.2 Accounting for human ratings using lexical indices

How well do these six dimensions account for perceptions of lexical richness? Whereas a handful of studies, which we will discuss shortly, used text-based indices to account for expert ratings of essay quality or of lexical proficiency, only one set of four related studies directly attempted to model non-experts' perceptions of the texts' lexical richness. In these studies, which were conducted by Jarvis (2013b, 2017) over the course of five years, different cohorts of 11 to 21 language teachers and linguistics students each rated the lexical richness of the same 50–60 English narrative retells on a 10-point scale. The texts were written by native speakers of English, Finnish, and Swedish, and varied in length between 24 and 578 words. The raters were not trained or given a rubric, but "lexical diversity" (Jarvis uses this term whereas we use "lexical richness") was defined for them as the variety of words in a text, and they were shown a sample text of average lexical diversity for reference. The results indicated, firstly, that the ratings were consistent across raters, with Cronbach's α for the three largest studies ranging between 0.90 and 0.96. Secondly, by scrambling the words in the texts, Jarvis (2017) ingeniously demonstrated that perceptions of lexical richness do not seem to be strongly affected by other factors related to writing quality or language proficiency (e.g., syntactic complexity and cohesion). Thirdly, the mean ratings per text could to a substantial degree be accounted for using indices that correspond to dimensions in Jarvis' framework: Jarvis (2013b) reports that six indices extracted from the texts accounted for about half of the variance in perceptions of lexical richness in the first—and least reliable—of the four studies ($R^2 = 0.48$).[3]

Jarvis' (2013b) study indicates that human ratings of lexical richness are predictable to some degree, but it could be fruitful to consider other operationalisations of the dimensions he proposed. Furthermore, it is not a priori clear if text-based metrics of

lexical richness should be linearly related to perceived lexical richness, so it may be worthwhile to experiment with statistical models that do not assume linear relationships between the indices and the perceptions. Finally, it is conceivable that the text-based indices stand in non-additive relationships to lexical richness perceptions (on the possibility of interacting effects in judgements of writing quality, see Jarvis et al., 2003), so it may also be worthwhile to allow for interaction effects in the models.

While Jarvis' (2013b, 2017) are the only studies that concerned lexical richness ratings by untrained raters, a number of studies sought to establish how well text-based indices predict expert ratings of overall text quality. Here, metrics associated with the variability (Crossley & McNamara, 2011; Engber, 1995; Grobe, 1981; Jarvis, 2002; Kuiken & Vedder, 2014; Malvern et al., 2004; McNamara et al., 2010), rarity (Crossley & McNamara, 2011; Guo, Crossley, & McNamara, 2013; Malvern et al., 2004; McNamara et al., 2010), and volume dimensions (Grobe, 1981; Jarvis et al., 2003; Nold & Freedman, 1977) helped to account for variance in the ratings. Additionally, variability, rarity, and volume metrics can be used as predictors for related constructs such as lexical proficiency and overall language proficiency (e.g., Crossley, Cobb, & McNamara, 2013; Crossley, Salsbury, & McNamara, 2011; Crossley, Salsbury, McNamara, & Jarvis, 2010; Kyle & Crossley, 2015; Laufer & Nation, 1995).

1.3 The present study

In the present study, we aim at furthering our understanding of text-based determinants of perceived lexical richness. That is, we follow Jarvis in treating lexical richness as a perceptual phenomenon: A text's perceived lexical richness is whatever a panel of uninstructed readers think it is. Our focus is primarily on the text-based indices' utility in predicting these perceptions.

The texts for which we collected these lexical richness perceptions stem from comparable French, German, and Portuguese corpora. While we will not address the question whether lexical richness perceptions can more accurately be predicted in one language compared to the others, analysing texts in three different languages in parallel offers the advantage that we can gauge to what extent the results (for instance, with respect to how particular text-based indices relate to perceptions of lexical richness or with respect to how strongly Jarvis' dimensional framework trades off predictive power for theoretical interpretability) are comparable across these three languages and may perhaps generalise to others. The issue of the robustness of the results with respect to the language that the texts were written in is also relevant in view of some more or less arbitrary language-specific decisions that researchers need to take when computing text-based indices. Examples of such decisions are whether to consider German *sie* 'she; they' and *Sie* 'you (formal)' as belonging to one, two or three different lemmata and whether to consider Portuguese fuses between prepositions and determiners (e.g., *neste* from *em* + *este* 'in this') as belonging to the preposition lemma (*em*), the determiner lemma (*este*), both lemmata (both *em* and *este*), or a *sui generis* lemma (*neste*).

An important aim of the present study concerns the trade-off between theoretical insight and predictive power when modelling perceptions of lexical richness. Typically, theoretical models are simplifying abstractions developed to render their subject matter more understandable. Such theoretical insight may come at the price of a loss in predictive power if the theoretical model disregards information that is available or makes simplifying assumptions about the relationship between the predictors and the outcome. Jarvis' (2013b) approach of conceptually deriving a limited number of dimensions relevant to perceived lexical richness and modelling human judgements using one (linear) predictor per dimension in a multiple regression model offers theoretical clarity. However, it would also be informative to assess how much predictive power is lost by limiting oneself to a few non-interacting, independent predictors in this way. Such insight would help researchers find out how strongly the theoretical model trades off data fit for conceptual understanding and may help them identify areas where the theoretical model can be improved upon (see Yarkoni & Westfall, 2017). To this end, we considered a wide array of predictors and allowed for the possibility that these have non-linear effects on human ratings and interact with one another. While multiple linear regression offers the advantage that its output is fairly interpretable, less transparent models and algorithms capable of identifying non-linear effects and interactions and dealing with a multitude of predictors (so-called *black boxes*) could yield greater predictive power. For this reason, the present study will explore, firstly, how well black-box algorithms can predict human ratings and, secondly, how much predictive power is lost by adopting more transparent models with fewer predictors selected for their fit with Jarvis' (2013a) framework.

The final aim of this study concerns the role of text length. Many measures of variability, by their design and due to the properties of languages, are affected by text length. This jeopardises comparisons of texts of different lengths on these measures. To solve these problems, researchers often constrain the text-length range in their sample (e.g., Crossley et al., 2010; Treffers-Daller, 2013; Treffers-Daller et al., 2016). However, comparing texts on such measures is not our goal; using these measures as predictors of perceived lexical richness is. If these perceptions are affected by the texts' length, then this is important to find out (see also Jarvis, 2013a). Relatedly, indices of variability are sometimes argued not to be valid for shorter texts. Koizumi (2012), for instance, recommends that the MTLD measure introduced by McCarthy and Jarvis (2010) should not be used for texts shorter than 100 tokens. The vast majority of the texts that we analyse, however, are shorter than that—which addresses a desideratum identified by McCarthy and Jarvis (2013) and Meara (2014). It is ultimately an empirical question how useful any text-based indices are for predicting lexical richness perceptions for short texts. Accordingly, the third goal of this study is to assess how strongly the predictability of human lexical richness ratings is compromised for very short texts.

To recapitulate, we ask how well text-based indices can predict uninstructed raters' perceptions of lexical richness for short French, German, and Portuguese texts if model interpretability is not an issue and how well more theoretically interpretable models

emulate this predictive power. Moreover, we ask if text-based indices are as useful for predicting the ratings of very short texts as they are for predicting the ratings of somewhat longer texts.

2. Method

We had a total of 3,060 French, German, and Portuguese texts written by 8- to 10-year-old children rated for their lexical richness by between three and 18 untrained native speakers each. From these texts, we extracted over 150 indices of word repetition, frequency etc., which we used to fit various statistical models in order to predict the average rating per text. To offset the danger of getting overly optimistic results inherent to exploratory analyses ('overfitting'), we cross-validated the models and, additionally, fitted them to independent hold-out sets.

All materials (texts, lemmatisation scripts, index-extraction scripts, Internet platform for collecting the ratings, datasets, and analysis scripts) are available from https://osf.io/vw4pc/. In order to keep this article readable, we only provide a summary of the study's methods as well as a couple of details that we deem crucial for understanding the results and judging the soundness of our conclusions. However, interested readers can refer to the project's technical report (Vanhove, 2018) for more details about all steps involved.

2.1 Texts

The texts were collected in a project that aimed to investigate the bilingual development of children with Portuguese as a heritage language (PHL) living in French- and German-speaking Switzerland (see Lambelet, Berthele, Desgrippes, Pestana, & Vanhove, 2017). 482 children wrote short argumentative and narrative texts at three points in time; they were on average 8;8 years old at the first data collection and 10;3 at the third. 114 of them were PHL speakers in French-speaking Switzerland; 119 were PHL speakers in the country's German-speaking part; the others were other pupils in French-speaking Switzerland (78), in German-speaking Switzerland (80), and in Portugal (91). PHL speakers wrote argumentative and narrative texts[4] in both Portuguese and French or German at each point in time, the rest only in their school language. Due to subject unavailability, some data are missing at each data collection.

The texts were corrected morphosyntactically, orthographically, and in terms of their punctuation. For instance, inappropriate or missing suffixes were corrected (e.g., German adjectives in the wrong case or missing French plural *-s*). Inappropriate lexical choices were not changed. Since we could not expect raters to judge the lexical richness of over 1,000 texts per language, we created twenty sets of 50–52 texts each for each language. Of the 3,760 texts that were written, we presented 3,060 texts to the raters. Texts consisting of fewer than 45 letters were not presented nor were texts without a single finite verb. Other than that, the selection was effected at random. The texts for each language were split up into twenty sets that were maximally similar in

terms of the number of texts written by children from the different regions, the number of argumentative vs. narrative texts, and the time at which the texts were written.

The texts retained for rating were short: the median text length was 37 tokens for French, 33 for German, and 39 for Portuguese. 90% of the French texts consisted of 81 tokens or fewer; for German, 66 tokens or fewer; and for Portuguese, 90 tokens or fewer. The full text length distributions are shown in Figure 1.

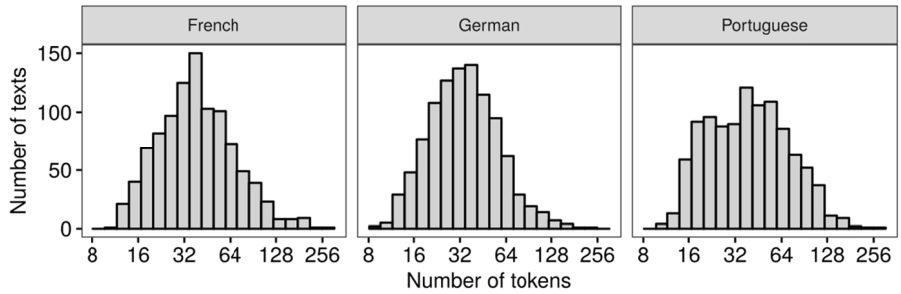

Figure 1: The number of tokens in the texts rated. Since these distributions are positively skewed, the data are plotted along a logarithmic *x*-axis.

2.2 Ratings

Raters. We recruited native speaker raters via social and professional networks and asked them to rate the lexical richness of the children's texts on an Internet platform. The raters were not paid for their participation.

Only raters who considered themselves native speakers of the respective language were included in the analysis. We only analysed data from raters who rated at least 48 texts.[5] The remaining ratings were checked for satisficing patterns (e.g., consistently providing the same rating), and data from four raters were discarded because of this. For the number of raters retained and further details, see Table 1.

Table 1. Description of the rater sample.

	French	German	Portuguese
Number of raters	146	322	106
Median age (years)	27	25	35
Percentage men	21	17	30
Percentage linguists	10	8	14
Percentage teachers	16	26	15
Percentage students	47	57	23

Procedure. The raters accessed an Internet platform where they first filled in a questionnaire. They were then asked to read the 50–52 texts and rate their lexical richness ('richness of the vocabulary used in the text') on a 9-point scale, the odd-numbered values of which were labelled (1 = very bad, 3 = fairly bad, 5 = average, 7 = fairly good, 9 = very good); see Figure 2. The instructions did not define lexical richness as we were interested in the raters' intuitive perception of lexical richness. This was also stressed in the instructions. The raters were told that the texts were written by children in primary school, but they were not told at which point in time the texts had been written nor that some texts were produced by children with a Portuguese background living in Switzerland.

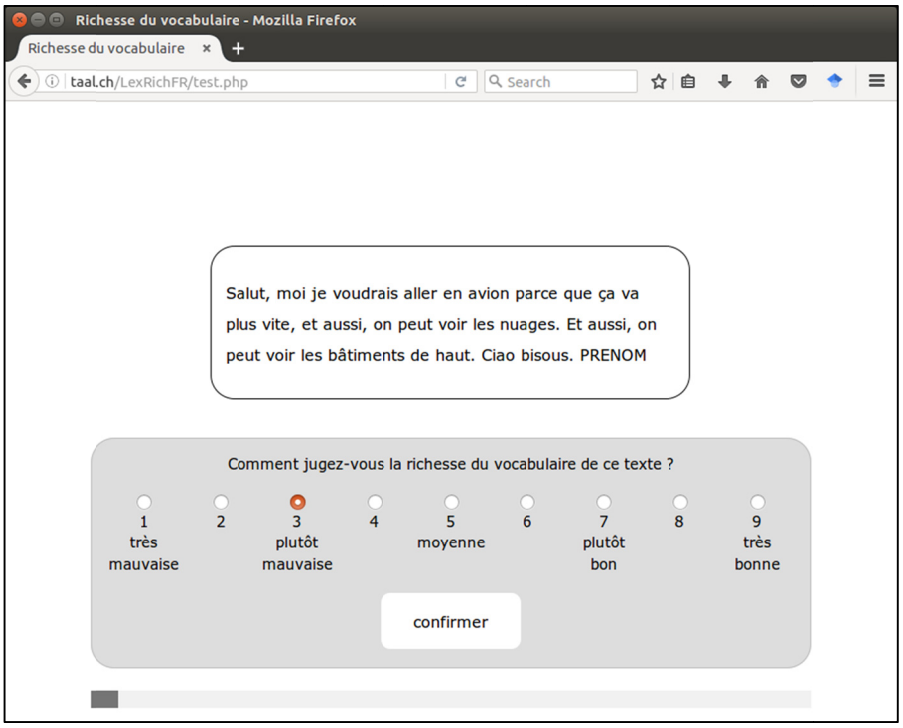

Figure 2. The French version of the rating interface.

The numbers of raters retained varied between the text sets, see Table 2. The raters saw all texts within the set assigned to them—presented in a new random order for each rater—and only those, with the exception of two additional texts at the beginning of the session which served to familiarise them with the task. These two texts were not included in the analyses.

Table 2. Number of texts and raters per set of texts.

	French	German	Portuguese
Number of text sets	20	20	20
Number of texts per set	50	51	52
Mean number of raters per set	7.3	16.1	5.3
Minimum number of raters per set	4	11	3
Maximum number of raters per set	9	18	6

2.3 Text-based indices

In order to compute the text-based indices, the texts were first tagged and lemmatised using the koRpus package (Michalke, 2017) for R (on the advantages of lemmatisation, see Treffers-Daller, 2013). This process involved a number of manual tweaks, which are fully documented in the project's technical report (Chapter 4). Once tagged and lemmatised, over 150 text-based indices were computed for each text. These are discussed in brief below; for a more detailed discussion, see the technical report (Chapters 5–7). They are organised here according to Jarvis' (2013a) six dimensions of perceived lexical richness.

Volume. Volume-related indices included the number of tokens, types, unique lemmata, part-of-speech-specific word counts, and the number of sentences.

Variability. The measures of variability computed included ten measures based solely on the number of types and tokens occurring in the texts, namely the type–token ratio (TTR), Guiraud's (1954) root-TTR, and eight other measures discussed by Tweedie and Baayen (1998). Lemma-based variants of these measures, e.g., Guiraud's root-TTR computed with respect to lemmata, were also computed. Type–token ratios were also computed split up by part of speech. Additionally, we computed Yule's (1944) *K*, three variants of Johnson's (1944) mean segmental type–token ratio (MSTTR), three variants of the HD-D value (McCarthy & Jarvis, 2007), four variants of Covington and McFall's (2010) moving-average type–token ratio (MATTR), and four variants of McCarthy and Jarvis' (2010) measure of textual lexical diversity (MTLD).

Evenness. Following Jarvis (2013b), evenness indices were computed by counting the number of tokens per type/lemma and calculating their standard deviation. Higher values reflect less evenness.

Rarity. Rarity measures were computed by looking up the frequencies of the words occurring in the text in frequency corpora. These measures were computed with respect to both the tokens and the lemmata in the text. The corpora used were Lexique 3 (New, Brysbaert, Veronis, & Pallier, 2007), SUBTLEX-DE (Brysbaert et al., 2011), and SUBTLEX-PT (Soares et al., 2015). We lemmatised SUBTLEX-DE and SUBTLEX-PT in order to compute lemma-based rarity measures; Lexique 3 already contains lemma information.

Four types of rarity measures were computed: frequency bands (cf. Laufer & Nation, 1995, e.g., Which proportion of the text do the 100 most frequent words in the language account for?), frequency summaries (cf. Kyle & Crossley, 2015, e.g., What is the mean corpus frequency of the words in the text?), frequency rank summaries (cf. Jarvis, 2013b, e.g., What is the mean corpus frequency of the words in the text?), and six variants each of Daller et al.'s (2003) 'advanced' type–token and Guiraud indices. The latter are computed like the ordinary TTR and Guiraud indices, but with only the number of 'advanced' (i.e., relatively rare) types in the numerator.

Disparity. No index of semantic disparity was computed. Formal disparity was not operationalised by Jarvis (2013b); our first attempt at a formal disparity measure was computed as follows. We computed string-edit distances between each type or unique lemma in the text and every other type or unique lemma in the text and then took the mean of all distances. The string-edit distances were computed using the Levenshtein algorithm, which computes the minimum number of operations (insertions, deletions, substitutions) required to transform one string into another. This operation cost was length-normalised (see the technical report, Chapter 7, for an example). Higher mean Levenshtein distances indicate more formal disparity in the text.

Dispersion. The dispersion index was computed following Jarvis (personal correspondence, August 2, 2017): For the *i*th word in the text ($i \in 1, 2, \ldots, n$), we looked up how often it occurred in the next k words (i.e., words i+1 through i+k). For all n words except the five most frequent in the language's frequency list, the number of 'close repeats' was summed and then divided by the total number of words. Six such dispersion indices were computed (varying k). Higher values reflect more 'clustericity' and hence less dispersion.

Miscellaneous. We computed measures of syntactic complexity (e.g., the number of conjunctions), lexical complexity (the mean number of letters per token), and lexical density (the proportion of tokens belonging to open word classes). A variable indicating whether the text was narrative or argumentative was also included as a predictor since this information was also available to the raters, but information about the children or the time when the text was written was not. Some of these measures (e.g., syntactic complexity) may seem unrelated to lexical richness proper but were included as they might conceivably affect untrained raters' perceptions.

2.4 Method of analysis

The goal of the analysis is to model the human ratings of the texts' lexical richness in terms of text-based properties. To this end, we used as the outcome variable the mean human rating per text. The text-based indices listed earlier served as potential predictors. Model fit was evaluated using the root mean square error (RMSE), that is, the square root of the mean squared discrepancy between the model's predictions ($\hat{y}_i$) and the actually observed values (y_i; i.e., RMSE = $\sqrt{\frac{1}{n}\sum(y_i - \hat{y}_i)^2}$). Since most other studies only report the coefficient of determination (R^2), this is also reported.[6]

Since it is not *a priori* clear which measures of variability, rarity etc. predict human ratings, this analysis is strongly exploratory. To offset the danger that extensive data exploration would lead to a model that tightly fits the present data but that does a poor job predicting similar but new data, we partitioned the data into a training and a test set, then applied various models and algorithms to the training set, used resampling techniques to adjudicate between different models, and, finally, used the models selected to predict the test set data. The caret package (Kuhn, 2017) for R served as the main computational tool.

Data partitioning. The training set consisted of a random selection of 16 out of 20 text sets per language; the test set consisted of the remaining 4 text sets. Since each text set was rated by a different panel of raters, the data in the test set are neither affected by the texts in the training set nor by the raters who rated the texts in the training set.

To the training set, a host of exploratory and modelling techniques were applied in order to find one or several statistical models with the greatest predictive power. The test set was not looked at until the very end of the analysis, at which point the final predictive models or algorithms constructed and selected on the basis of the training sets were applied to the test set in order to assess their predictive accuracy for *different texts judged by different rater panels*. At no point were the test data used to construct or re-estimate the models or algorithms. Moreover, the models were not respecified after we learnt about their performance in the test set.

Exploratory and modelling techniques. We applied a series of models and algorithms ('models' for short), many of which capable of dealing with a fairly large number of predictors and of capitalising on non-linearities and interaction effects. In addition to multiple linear regression (with and without prior principal component analysis), these were robust regression, ridge regression, elastic net, multivariate adaptive regression splines, partial least squares regression, *k*-nearest neighbours, regression trees, random forests (both CART and conditional inference-based), support vector machines, stochastic gradient boosting, and Cubist. A technique known as 'stacking', in which predictions from multiple models are combined and which may yield further improvements in predictive accuracy (see Breiman, 1996), was also applied; fourteen models per language were stacked. We do not discuss the architecture of these different models but instead refer to Chapters 5 through 8 in Kuhn and Johnson (2013). What is important here is that by applying these different models we were able to gauge how well human ratings of lexical richness could be predicted on the basis of text-based properties if interpretability were not an issue.

Many of these models are known as 'black boxes', that is, the precise way in which they relate a set of input values to a predicted outcome can be difficult to understand. Indeed, even for linear regression models, it can be difficult to assess the independent effect of each predictor on the outcome: a model coefficient might tell you how the outcome is expected to change when varying the TTR while keeping the numbers of types and tokens constant, but this is obviously impossible to do. Moreover, models with substantial different architectures often had similar predictive power.

While this underscores that when it comes to predicting outcomes on the basis of moderately rich predictor data, there are many ways to skin a cat (the 'Rashomon effect', see Breiman, 2001), it further compounds their lack of interpretability. Thus, we also fitted more transparent models whose construction was guided by Jarvis' (2013a) 6-dimensional theoretical framework (a 6-predictor model) to see how well the black boxes' predictive power could be emulated using only a handful of predictors.

When exploring the training data, we noticed that it may be possible to achieve reasonable predictive accuracy using a simpler model still—simpler both in terms of how the predictor values could be computed and in terms of how easily these predictor values could be translated into a predicted lexical richness perception rating. We therefore also included in the Results section the results of this 'single-predictor approach'.

For reference, the performance of an intercept-only regression model (without any predictors) is also reported; such a model would predict all test data to be equal to the mean of the training data. (This is the $\hat{y}_0$ value mentioned in Note 6.)

Cross-validation. When trying out different models on the training data, we used cross-validation to estimate how well the models would work for new data. This was done to ensure that overzealous data exploration and model fine-tuning would not result in a model that fits the training data well but stands little chance of predicting the test data (see Kuhn & Johnson, 2013; Yarkoni & Westfall, 2017). In cross-validation, the training data is split up into a number (k) of folds, and models are fitted on $k - 1$ folds and then used to predict the outcome in the remaining fold. This process is repeated k times, each time leaving out a different fold, which yields k estimates of the models' predictive accuracy on unseen data (viz., the root mean square error, RMSE) that can then be averaged. To account for the dependency structure in the data (some mean ratings are based on ratings by the same judges), block cross-validation was used: rather than constructing the folds randomly, each of the sixteen text sets per language in the training data was used 15 times in its entirety for training and once for prediction. This way, the texts in each predicted fold were all rated by different raters from the texts in the other 15 folds. Figure 3 illustrates the principles behind the partitioning of the data and block cross-validation.

Predictor selection and transformation. Unsurprisingly, many predictor variables turned out to be strongly linearly or non-linearly correlated with each other. While collinearity does not dramatically affect the performance of predictive models, (near-) perfect correlations between predictors were resolved by dropping some of the offending predictors. The choice about which variable to drop was made mostly on a conceptual basis. For instance, when type- and lemma-based variables were nearly perfectly correlated, the lemma-based variable was retained. In total, 111 predictors were retained for each language. All of these were fed to the black-box models.

Many of the predictors were severely right-skewed, so that a Yeo–Johnson transformation (a generalisation of the Box–Cox family of transformations that can

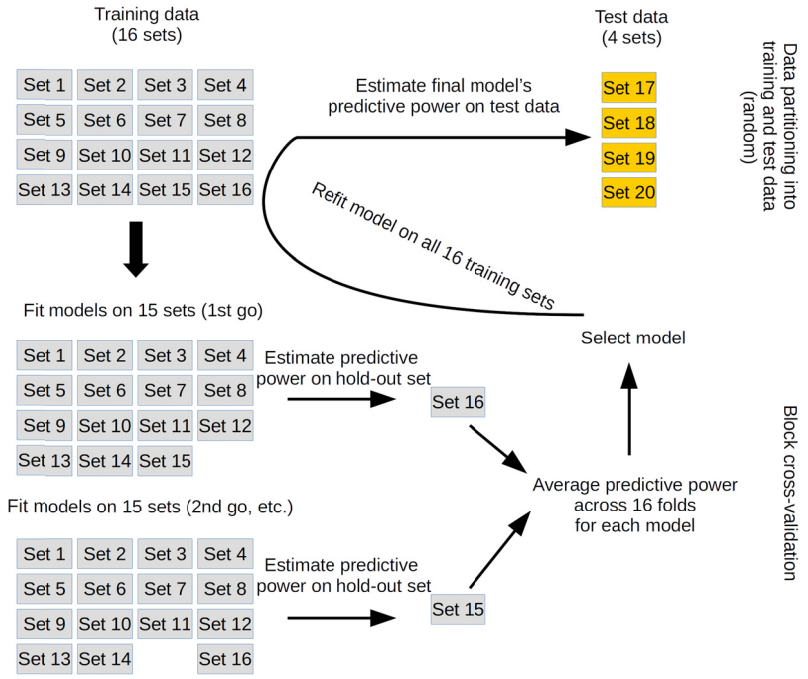

Figure 3. Illustration of how the data were partitioned into a training and a test set and of how block cross-validation works. Only two iterations of block cross-validation are shown; in reality, sixteen took place for each model. Each 'Set' refers to 50-52 texts that were rated by a panel of judges. The panels of judges for different sets did not overlap.

accommodate zeroes and negative values; Yeo & Johnson, 2000) was applied to the entire predictor set; the form of the transformation was determined on the basis of the training set only. The predictors were subsequently centred at their training set means and scaled using their training set standard deviations.

3. Results

3.1 Systematicity and reliability

A reliability analysis suggests that the judgements of untrained raters are to a large extent systematic. For each rater panel, that is, for each set of texts in each language, we computed the reliability of the mean ratings in terms of the ICC(2,*k*) reliability coefficient (where *k* is the number of raters) as recommended by Shrout and Fleiss (1979) using the psych package for R (Revelle, 2017). Unlike Cronbach's α, this coefficient assumes that the raters in the study are but a subset of the population of raters one is actually interested in and wishes to generalise to. For French, the twenty ICC(2,*k*) coefficients averaged 0.79; for German, 0.90; for Portuguese, 0.71. The higher

reliability for German primarily reflects the fact that the mean ratings were based on more individual ratings. These reliability estimates are clearly lower than Jarvis (2017) reported; we will return to these differences in the Discussion.

The reliability of the outcome variable (mean perceived lexical richness) effectively puts a ceiling on how well a model can be expected to predict new data, so that the models for French and Portuguese can be expected to have poorer predictive power than the one for German. Hence, we do not compare models' predictive power between languages but only within each language.

3.2 Overall predictive accuracy

Black box approach. Several black-box models had a similar predictive accuracy in cross-validation (see technical report, Chapters 11–13). Model stacking, in which the outcomes predicted by several individual models were used as predictors in a second-level model, tended to yield slightly improved predictions in cross-validation and was consequently used to gauge how well lexical richness ratings can be predicted if model interpretability is of little importance.

Figure 4 shows the predictive accuracy of the black-box approach in cross-validation and in the independent test sets. Relative to the reference model containing no predictors, this approach yields a substantial improvement in predictive accuracy: while the average error (roughly speaking) of the oversimplistic models hovers around 1.2 points on the 9-point scale for French and German and around 1.3–1.5 points for Portuguese, it is reduced by some 0.31–0.36 points for French and some 0.45–0.48 points for German and Portuguese. For the sake of completeness, Figure 4 also shows the coefficients of determination of the black-box approach, but see Note 6 about the use of R^2 for evaluating model fit.[7]

Six-dimension approach. While we can predict human ratings with some degree of accuracy by taking a black-box approach, this offers little in the way of theoretical interpretability. We also modelled the human ratings in smaller, more interpretable models in which each of the six dimensions identified by Jarvis (2013a, 2013a) was represented by a single predictor variable. For volume, the number of tokens was chosen as the most straightforward operationalisation. For variability, the MTLD was chosen as it is in principle orthogonal to text length; its TTR threshold was set to 0.83.

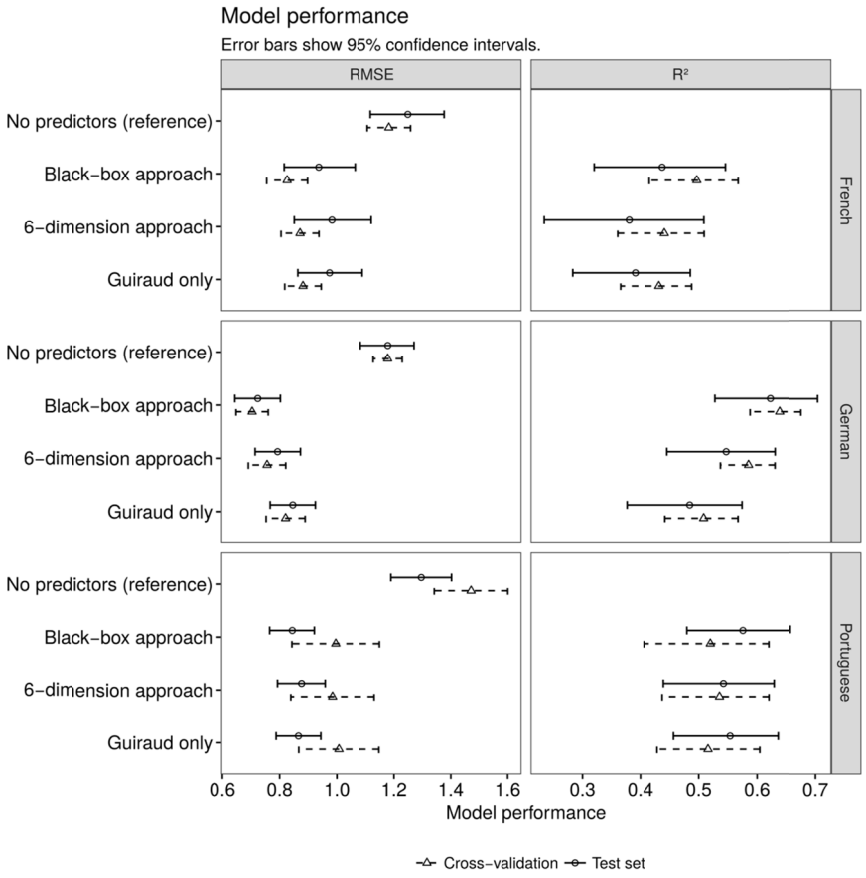

Figure 4. The root mean squared errors (RMSE) and the coefficients of determination (R^2) of the different approaches to predicting average lexical richness ratings in both cross-validation and independent test sets. Lower RMSEs indicate greater predictive accuracy; higher R^2 values indicate a stronger decrease in the residual sum of squares of the predictive model relative to the reference model (see Note 6). No R^2 was computed for the reference model as it is 0 by definition.

For rarity, the mean Zipf score of the unique lemmata occurring in the text was used, though other predictors yielded similar results in cross-validation.[8] For evenness, disparity, and dispersion, the respective lemma-based operationalisations were chosen. The predictor variables were transformed as needed to reduce the skew in their distribution, but the evenness operationalisation in particular was collinear with other predictors; see Table 3 for a summary. Their relation to the mean ratings was then modelled in generalised additive models (GAMs) using the mgcv package (Wood, 2017) for R. GAMs do not presuppose that the predictors are linearly related to the

outcome. Instead, they allow these relationships to be modelled nonlinearly, with the degree of nonlinearity being estimated from the data (for an introduction to GAMs, see Clark, 2016). The predictive power of these models was again assessed using block cross-validation.

Table 3. Six dimensions of lexical richness and the predictor variables representing them. Four of the six predictors were logarithmically or square-root transformed in order to reduce the positive skew. The choice between these two transformations was made on the basis of the training data only. For more information about these predictors, their intercorrelations and their relationship to the ratings, see the technical report (Vanhove, 2018), Figures 11.21, 12.21 and 13.21.

Dimension	Predictor	Comments
Volume	Number of tokens	Logarithmically transformed
Variability	MTLD with TTR threshold 0.83	Logarithmically transformed
Evenness	Standard deviation of tokens per lemma	Square-root transformed; strongly correlated with volume ($0.78 < r < 0.80$) and dispersion ($0.62 < r < 0.70$)
Rarity	Mean Zipf of unique lemmata	
Disparity (formal)	Mean Levenshtein distance between unique lemmata	
Dispersion	Dispersion index for lemmata, $k = 20$	Square-root transformed; collinear with variability ($-0.66 < r < -0.61$)

Figure 4 shows these models' predictive usefulness ('6-dimension approach'). In terms of their RMSE, these models only differ slightly from the black-box models, with the latter at best outperforming them by 0.07 points on the 9-point scale. While some of these differences are significant (see technical report, Chapters 11–14), a difference of 0.07 on a 9-point scale is obviously minute, especially considering that this approach used 105 fewer variables than did the slightly more powerful one.

Figure 5 shows the partial effects of the six predictors on the mean rating as estimated using the entire training data set. While there are, of course, some differences between the models for the three languages, the similarities are striking. First, the 'volume' predictor is the strongest of the six predictors, with longer texts receiving better ratings. Second, the 'variability' predictor has a consistent effect in the expected direction, but its effect is clearly small. Third, other things equal, texts in which the standard deviation of the number of tokens per lemma are large, i.e., which show a less even spread of tokens over lemmata, are rated as lexically poorer. This is also what Jarvis (2013a) expected, though ideally, this finding should be replicated using a measure of evenness that is less strongly correlated with the other predictors. Fourth, texts with less frequent words are rated as lexically richer in all three languages, which

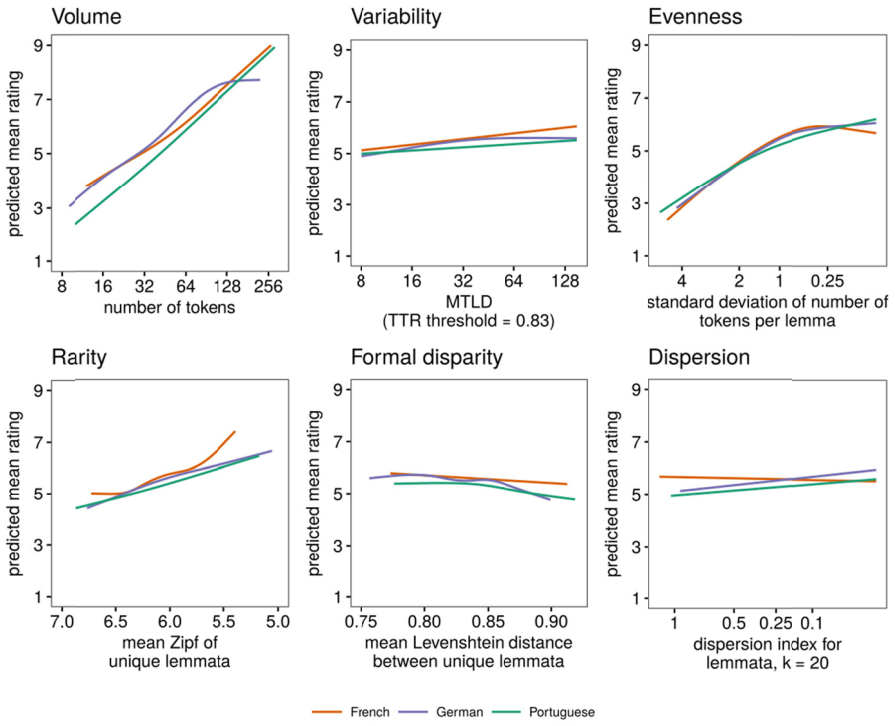

Figure 5. Partial effects of the six predictors fitted in generalized additive models on the basis of the training sets. The *x*-axes have been arranged such that values to the right indicate more volume, variability, etc. The lines in each panel show how the mean ratings predicted by the French, German, and Portuguese models vary when the predictor in question was allowed to vary along its range in the training set. The five predictors not shown in any given panel were fixed at their training set mean when plotting. Language-specific versions of this figure with confidence bands for the partial effects are available in the technical report (Vanhove, 2018; Sections 11.8, 12.8 and 13.8).

again meshes with Jarvis' expectations. Fifth, formal disparity does not seem to affect the ratings much; if at all, the effect runs counter to Jarvis' expectations, with texts exhibiting greater formal disparity being rated as slightly lexically poorer. Sixth, dispersion does not seem to strongly affect lexical richness ratings; if at all the effect runs counter to Jarvis' expectations, with texts in which lemmata are more dispersed being rated as lexically poorer in German and Portuguese.

Single-predictor approach. In the third and last modelling approach, we wanted to explore how much predictive power is lost, relative to the first two approaches, by predicting lexical richness ratings in terms of a single, easily computable and applicable model. Variable importance measures (see Breiman, 2001; Kuhn & Johnson, 2013, Chapter 18) identified Guiraud's index as the single most important predictor of lexical richness judgements in the black-box models (see technical report, Sections 11.7, 12.7

and 13.7). This predictor is straightforward to compute as $\frac{\text{number of unique lemmata}}{\sqrt{\text{number of tokens}}}$ and so does not require the analyst to interface the texts with an external frequency list, to distinguish the words in the texts by part of speech, etc. Guiraud's index was not included in the 6-dimension model because of its strong correlation with text length ($0.76 < r < 0.85$ in the training sets).

As Figure 4 shows ('Guiraud only'), linear regression models with this single predictor did a respectable job relative to the black-box and 6-dimension models: while the RMSE of the Guiraud-only models was between 0.01 and 0.12 points higher than that of the black-box models, and between -0.01 and 0.06 higher than that of the 6-dimension models, such differences are rather modest considering that the data are on a 9-point scale. In sum, while the more complex models may be preferable for predictive and theoretical reasons, these simple models may be more than serviceable when a rough-and-ready gauge of a text's perceived lexical richness is desired. Equations (1) to (3) provide the regression equations.[9]

$$(1)\quad \text{predicted mean rating (French)} = 1.44 + 0.81 \times \frac{\text{number of lemmata}}{\sqrt{\text{number of tokens}}}$$

$$(2)\quad \text{predicted mean rating (German)} = 0.77 + 1.03 \times \frac{\text{number of lemmata}}{\sqrt{\text{number of tokens}}}$$

$$(3)\quad \text{predicted mean rating (Portuguese)} = 0.11 + 1.04 \times \frac{\text{number of lemmata}}{\sqrt{\text{number of tokens}}}$$

3.3 Predictive accuracy and text length

Finally, we assessed whether the usefulness of text-based indices for predicting the ratings is compromised for very short texts. To this end, we plotted the absolute prediction errors, that is, the discrepancies between the actual mean lexical richness ratings and the predicted mean lexical richness ratings, from the black-box approach against the length of the texts; see Figure 6. If predictability were compromised for very short texts, then the average prediction error—highlighted by the scatterplot smoothers—would be consistently larger for short compared to longer texts. Clearly, this is not the case, and we conclude that, within this set of fairly short texts, the predictability of lexical richness ratings is not compromised for the shortest of texts

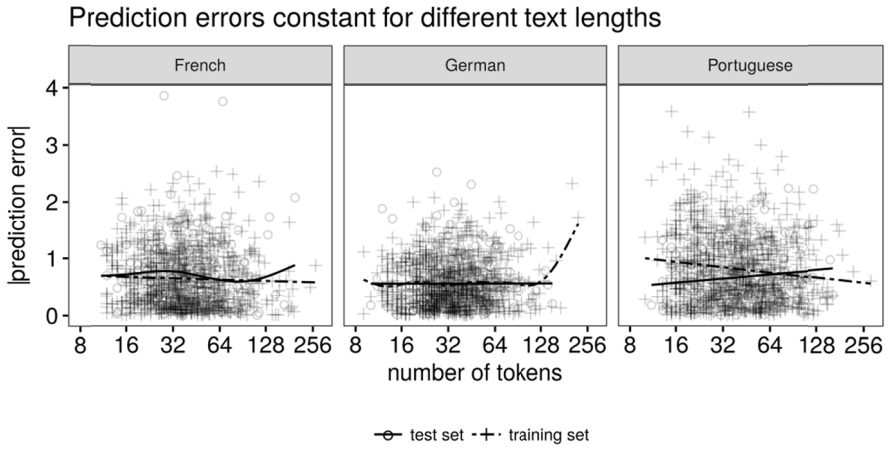

Figure 6. The absolute prediction error for each text plotted against the text's length; the average absolute prediction error is highlighted by scatterplot smoothers and does not vary consistently according to text length. The predictor errors are derived from the black-box approach. For the training set, the prediction errors are based on out-of-fold predictions (i.e., on the cross-validation run in which the datum was not used to fit the model).

4. Discussion

The overarching goal of this study was to elucidate the predictability of the lexical richness of short texts as perceived by untrained raters. To sum up our results, we found (a) that the lexical richness ratings of untrained raters exhibited a respectable degree of inter-rater reliability, though not to the extent found in a series of related studies (Jarvis, 2017); (b) that these ratings are partly predictable on the basis of text-based indices; (c) that only little predictive power is lost by taking a theory-driven approach (as suggested by Jarvis, 2013a) as opposed to a 'kitchen-sink' tack; (d) that only little predictive power is lost by considering just a single, easy-to-compute predictor, viz., Guiraud's root-TTR; and (e) that predictability is not worse for the shortest of texts within this set of fairly short texts.

4.1 Systematicity and validity

The reliability coefficients indicate that the lexical richness ratings are clearly non-random. Nonetheless, their reliabilities may seem modest compared to the high reliabilities ($\alpha > 0.90$) that Jarvis (2017) reports. We see four reasons for this difference.

First, we had fewer raters per text (3-18) than did Jarvis (20-21), and Cronbach's α and ICC(2,*k*) values are naturally larger the more individual ratings go into the mean ratings. In practical terms, future studies of lexical richness may wish to employ more raters per text than we did, at the cost of analyzing fewer texts.

Second, our reliability estimate (ICC(2,*k*)) assumes that the raters who rated a given set of texts form but a subset of the raters in whose judgements one is actually interested in ('treating raters as random'). That is, it estimates how similar the mean ratings assigned by *different*, equal-sized panels of raters would be. Cronbach's α, by contrast, assumes that only the present panel of raters is of interest ('treating raters as fixed'). Since we were interested in how well the present findings generalise to both new texts new raters the assumption underlying the ICC(2,*k*) seemed better suited to our purpose. For comparison, the mean reliabilities assuming raters as a fixed effect are 0.84, 0.93, and 0.81, for French, German, and Portuguese, respectively.

Third, the instructions we gave to our raters were even less elaborate than Jarvis did to his. Specifically, we asked our raters to rate the vocabulary richness of the texts, but unlike Jarvis, we did not define this, instead preferring to let raters use whichever concept the term evoked for them. Moreover, we did not provide them with a benchmark text representing average lexical richness.

Fourth, Jarvis' raters were students in his linguistics class, and he was able to incentivise them to rate the texts consistently. In contrast, our rater sample consisted of a more varied mix of raters who logged on to an Internet platform and to whom we did not offer any incentives.

All things considered, our results indicate that Jarvis' (2017) finding that lexical richness ratings by untrained raters are consistent also applies to shorter texts, thereby confirming Meara's (2014) intuition that people can make reliable lexical richness judgements on the basis of little material and with little guidance.

That said, our decision not to provide raters with a more elaborate definition of lexical richness than "the richness of the vocabulary used in the text" may conceivably have led at least some of them to judge the quality of the texts in more general terms than specifically their lexical richness. To account for this possibility, future studies may wish to experiment with varying the explicitness of the instructions to the raters or by scrambling the words in the texts such that non-lexical markers of writing quality (e.g., syntactic complexity, structure) cannot affect the raters' perceptions (cf. Jarvis, 2017).

4.2 Predictability and interpretability

In addition to being systematic, the lexical diversity ratings are predictable: up to a certain degree, it can be forecast how new panels of raters would judge the lexical richness of new texts. This predictability is not compromised for the shortest of texts.

What is more, fairly interpretable models based on Jarvis' (2013a) theoretical framework which contain only six, non-interacting predictors do not compromise predictability. These models identify 'volume' as the most important predictor of lexical richness, which replicates a finding by Jarvis (2013b). 'Volume' is followed by 'evenness' and 'rarity', which both have effects in the direction expected. In contrast to these three predictors, the MTLD measure of 'variability', our first attempt at a 'formal disparity' measure, and the 'dispersion' measure proposed by Jarvis (personal correspondence, August 2, 2017), did not turn out to be important predictors.

However, some of these measures express the same information to some degree. For instance, short texts tend to have low type–token ratios, which translates into high evenness and high dispersion. The development of operationalisations of these dimensions that are not so inextricably intertwined would facilitate the interpretation of the results of future studies.

A further observation was that simple models containing only a single predictor, Guiraud's root-TTR, yielded fairly accurate predictions. This index is easy to compute, and the model predictions easy to calculate. For all its predictive prowess, however, the relationship between Guiraud's root-TTR and the ratings is difficult to interpret as this variable is strongly correlated with log-transformed text length ($0.76 < r < 0.85$ in the training sets) and fairly strongly with other indices of variability as well (e.g., with the log-transformed MTLD-0.83 measure; $0.38 < r < 0.42$). Consequently, Guiraud's index is best considered an amalgamation of 'volume' and 'variability' in these texts, making it difficult to determine just how much predictive power it owes to which aspect.

4.3 Improving predictability

While different models could predict lexical richness ratings to a certain degree, the prediction is of course far from perfect. Apart from ensuring that their richness ratings are highly reliable by recruiting twenty or so raters per text, researchers seeking greater predictive accuracy may wish to explore different ways of operationalising the constructs suggested by Jarvis (2013a, 2017). For instance, Jarvis (2017) currently defines 'specialness' in terms of how much individual words contribute to a text's meaning, and it may be worthwhile to try to operationalise this in terms of the words' predictability given the context. Moreover, while Jarvis (2013b) operationalised 'disparity' in semantic terms, we only measured it in formal terms, and this operationalisation is to be understood as a first attempt in want of refinement.

Consequently, we intend this report also as an invitation to other researchers to test the limits of the predictability of lexical richness ratings by extracting different indices from the texts and using them as predictors in predictive models—or indeed by collecting additional ratings. By doubly validating these models (cross-validation and validation on an independent test set), the risk of overfitting typically associated with such data exploration should be largely nullified. We particularly wish to stress this latter point: Research on modelling perceptions of lexical richness/diversity is still in its infancy, and we think it is of the utmost importance that the first studies do not raise unrealistic expectations by reporting impressive-looking findings with little chance of generalising to new texts and new raters.

5. Conclusions

Lexical richness ratings are both systematic and up to a degree predictable, even for short texts. Little predictive accuracy is lost by adopting theoretically rooted predictive models rather than opaque algorithms. We suggested some avenues for further

improvements to the theoretical and statistical models and stressed the need for validating these models in order to estimate how applicable they are to both different texts and different raters.

Notes

1. Some researchers use the term *lexical diversity* instead of *lexical richness* to refer to the quantity, range, and variety of the vocabulary used in a text (e.g., Jarvis, 2013a), whereas others use the term *lexical diversity* to refer to the variety aspect only and use *lexical richness* as the superordinate term (e.g., Engber, 1995; Read, 2000). In the rating study discussed in this article, we asked the raters to judge the texts' lexical richness, so it is this term we use throughout the article.
2. Currently, Jarvis (2017) distinguishes between seven dimensions: volume, abundance, variety, evenness, dispersion, specialness, and disparity. The present study was carried out before this new proposal was published, so we mostly refer to the old one.
3. Jarvis (2017) mentions that more refined indices could account for 89% of the variance in the last—and most reliable—studies, but further details, such as the way in which the indices were computed, how the data were modelled, and how the danger of overfitting was addressed, are not provided.
4. For the argumentative texts, the children wrote a letter to their aunt or godmother in which they needed to convince a family member to go on a holiday to either the seaside or into the mountains (Portuguese) or to go on a holiday by plane or by car (French and German). For the narrative texts, they needed to relate what happened during their last holidays (Portuguese) or on the last school trip (French and German).
5. Several raters quit after having rated only a handful of texts. For a couple of raters who persisted until the end, a handful of ratings were not logged due to technical glitches, hence the selection criterion of 48 out of 50–52 texts.
6. We use the RMSE as it expresses directly how well the model predictions correspond to the observed values. The problem with R^2 is that there exist different formulae for computing R^2 (see Kvålseth, 1985). For ordinary regression models, these all yield the same result. However, when the model is used to predict observations that were not used when fitting the model, they do not. One popular method for computing R^2, namely squaring the correlation between the predicted and observed values, is particularly problematic, since the correlation between the predicted and observed values can be excellent even if the former correspond poorly to the latter (e.g., the values 1, 2, 3 correlate perfectly with the values 2000, 4000, 6000 but correspond poorly to them). The R^2 values in this article were therefore calculated as the proportional reduction in the residual sum of squares relative to a baseline model without any predictors: $R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \hat{y}_0)^{2'}}$ where y_i is a text's perceived lexical richness, $\hat{y}_i$ its predicted perceived lexical richness by the

predictive model, and $\hat{y}_0$ its 'predicted' perceived lexical richness by a model without any predictors (i.e., an intercept-only model).

7. To respond to a reviewer inquiry, we do not report *p*-values for four reasons. First, these are not intended for assessing a model's predictive strength. Second, we do not test null hypotheses in this study. Third, even if we did, many of the models we used do not output *p*-values. And fourth, even if they did, it would not be clear how we would have to adjust them to take into account the strong exploratory component in our analyses (see Altman & Krzywinski, 2017). Similarly, we do not provide lengthy tables of model coefficients for the black-box models. This is because some of these models (e.g., the tree-based and neighbour-based models) do not have model coefficients as one may know them from regression analyses, whereas the models that do have model coefficients have too many of them to be interpreted. For the generalised additive models we used in the 6-dimension approach further in the main text, we do not report estimated model coefficients because these cannot be meaningfully interpreted either; generalised additive models need to be visualised to be interpreted (cf. Figure 5).
8. The Zipf scale was proposed by van Heuven, Mandera, Keuleers, and Brysbaert (2014) as a frequency scale that captures language users' perception of the relative frequency of words.
9. The performance of linear models with a type-based Guiraud's index (which is even easier to compute as it does not require a lemmatisation of the texts) is virtually indistinguishable from those with a lemma-based Guiraud's index, but the regression equations do differ slightly; see technical report.

Acknowledgements

This study was funded by the Research Centre on Multilingualism (Fribourg). We thank Judith Berger, Katharina Karger, Meik Michalke, Carlos Pestana, Catia da Silva Parente, Fabio Soares, Carina Steiner, and Isabelle Udry for their help at various stages in this project.

References

Altman, N., & Krzywinski, M. (2017). Points of significance: *p* values and the search for significance. *Nature Methods*, *14*, 3–4. doi:10.1038/nmeth.4120

Breiman, L. (1996). Stacked regressions. *Machine Learning*, *24*, 49–64. doi:10.1007/BF00117832

Breiman, L. (2001). Statistical modeling: The two cultures. *Statistical Science*, *16*(3), 199–231. doi:10.1214/ss/1009213726

Brysbaert, M., Buchmeier, M., Conrad, M., Jacobs, A. M., Bölte, J., & Böhl, A. (2011). The word frequency effect: A review of recent developments and implications for the choice of frequency estimates in German. *Experimental Psychology*, *58*, 412–424. doi:10.1027/1618-3169/a000123

Clark, M. (2016, June). Generalized additive models. Retrieved from https://m-clark.github.io/docs/GAM.html

Covington, M. A., & McFall, J. D. (2010). Cutting the Gordian knot: The moving-average type–token ratio (MATTR). *Journal of Quantitative Linguistics*, *17*(2). doi:10.1080/09296171003643098

Crossley, S. A., & McNamara, D. S. (2011). Understanding expert ratings of essay quality: Coh-Metrix analyses of first and second language writing. *International Journal of Continuing Engineering Education and Life Long Learning*, *21*(2-3), 170–191. doi:10.1504/ IJCEELL.2011.040197

Crossley, S. A., Cobb, T., & McNamara, D. S. (2013). Comparing count-based and band-based indices of word frequency: Implications for active vocabulary research and pedagogical applications. *System*, *41*, 965–981. doi:10.1016/j.system.2013.08.002

Crossley, S. A., Salsbury, T., & McNamara, D. S. (2011). Predicting the proficiency level of language learners using lexical indices. *Language Testing*, *29*(2), 243–263. doi:10.1177/0265532211419331

Crossley, S. A., Salsbury, T., McNamara, D. S., & Jarvis, S. (2010). Predicting lexical proficiency in language learner texts using computational indices. *Language Testing*, *28*(4), 561–580. doi:10.1177/0265532210378031

Daller, H., van Hout, R., & Treffers-Daller, J. (2003). Lexical richness in the spontaneous speech of bilinguals. *Applied Linguistics*, *24*(2), 197–222. doi:10.1093/applin/24.2.197

Engber, C. A. (1995). The relationship of lexical proficiency to the quality of ESL compositions. *Journal of Second Language Writing*, *4*(2), 139–155. doi:10.1016/1060-3743(95)90004-7

Grobe, C. (1981). Syntactic maturity, mechanics, and vocabulary as predictors of quality ratings. *Research in the Teaching of English*, *15*(1), 75–85.

Guiraud, P. (1954). *Les charactères statistiques du vocabulaire. Essai de méthodologie*. Paris: Presses Universitaires de France.

Guo, L., Crossley, S. A., & McNamara, D. S. (2013). Predicting human judgments of essay quality in both integrated and independent second language writing samples: A comparison study. *Assessing Writing*, *18*, 218–238. doi:10.1016/j.asw.2013.05.002

Jarvis, S. (2002). Short texts, best-fitting curves and new measures of lexical diversity. *Language Testing*, *19*(1), 57–84. doi:10.1191/0265532202lt220oa

Jarvis, S. (2013a). Capturing the diversity in lexical diversity. *Language Learning*, *63*(Supplement 1), 87–106. doi:10.1111/j.1467-9922.2012.00739.x

Jarvis, S. (2013b). Defining and measuring lexical diversity. In S. Jarvis & M. Daller (Eds.), *Vocabulary knowledge: Human ratings and automated measures* (pp. 13–43). Amsterdam: John Benjamins. doi:10.1075/sibil.47

Jarvis, S. (2017). Grounding lexical diversity in human judgments. *Language Testing*, *34*(4), 537–553. doi:10.1177/0265532217710632

Jarvis, S., Grant, L., Bikowski, D., & Ferris, D. (2003). Exploring multiple profiles of highly rated learner compositions. *Journal of Second Language Writing*, *12*(4), 377–403. doi:10.1016/j.jslw.2003.09.001

Johnson, W. (1944). Studies in language behavior: I. A program of research. *Psychological Monographs*, *56*, 1–15.

Koizumi, R. (2012). Relationships between text length and lexical diversity measures: Can we use short texts of less than 100 tokens? *Vocabulary Learning and Instruction*, *1*(1), 60–69. doi:10.7820/vli.v01.1.koizumi

Kuhn, M. (2017). *caret: Classification and regression training*. R package, version 6.0-7.6. Retrieved from https://github.com/topepo/caret/

Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling*. New York: Springer. doi:10.1007/978-1-4614-6849-3

Kuiken, F., & Vedder, I. (2014). Rating written performance: What do raters do and why? *Language Testing*, *31*(3), 329–348. doi:10.1177/0265532214526174

Kvålseth, T. O. (1985). Cautionary note about R^2. *The American Statistician*, *4*(1). doi:10.2307/2683704

Kyle, K., & Crossley, S. A. (2015). Automatically assessing lexical sophistication: Indices, tools, findings, and application. *TESOL Quarterly*, *49*(4). doi:10.1002/tesq.194

Lambelet, A., Berthele, R., Desgrippes, M., Pestana, C., & Vanhove, J. (2017). Testing interdependence in Portuguese heritage speakers in Switzerland: The HELASCOT project. In R. Berthele & A. Lambelet (Eds.), *Heritage and school language literacy development in migrant children: Interdependence or independence?*, pp. 58-82. Bristol, UK: Multilingual Matters.

Laufer, B., & Nation, P. (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied Linguistics*, *16*(3), 307–322. doi:10.1093/applin/16.3.307

Malvern, D., Richards, B., Chipere, N., & Durán, P. (2004). *Lexical diversity and language development: Quantification and assessment*. Basingstoke, UK: Palgrave Macmillan. doi:10.1007/978-0-230-51180-4

McCarthy, P. M., & Jarvis, S. (2007). vocd: A theoretical and empirical evaluation. *Language Testing*, *24*(4), 459–488. doi:10.1177/0265532207080767

McCarthy, P. M., & Jarvis, S. (2010). MTLD, voc-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, *42*(2), 381–392. doi:10.3758/BRM.42.2.381

McCarthy, P. M., & Jarvis, S. (2013). From intrinsic to extrinsic issues of lexical diversity assessment: An ecological validation study. In S. Jarvis & M. Daller (Eds.), *Vocabulary knowledge: Human ratings and automated measures* (pp. 45–77). Amsterdam: John Benjamins. doi:10.1075/sibil.47

McNamara, D. S., Crossley, S. A., & McCarthy, P. M. (2010). Linguistic features of writing quality. *Written Communication*, *27*(1), 57–86. doi:10.1177/0741088309351547

Meara, P. (2014). [Review of the book *vocabulary knowledge: Human ratings and automated measures*, by Scott Jarvis and Michael Daller (eds.)]. *International Journal of Applied Linguistics*, *24*(3), 418–421. doi:10.1111/ijal.12086

Michalke, M. (2017). *koRpus: An R package for text analysis*. R package, version 0.10-1. Retrieved from http://reaktanz.de/?c=hacking&s=koRpus

New, B., Brysbaert, M., Veronis, J., & Pallier, C. (2007). The use of film subtitles to estimate word frequencies. *Applied Psycholinguistics*, *28*, 661–677. doi:10.1017/S014271640707035X

Nold, E. W., & Freedman, S. W. (1977). An analysis of readers' responses to essays. *Research in the Teaching of English*, *11*(2), 164–174.

Read, J. (2000). *Assessing vocabulary*. Cambridge, UK: Cambridge University Press.

Revelle, W. (2017). *psych: Procedures for psychological, psychometric, and personality research*. R package, version 1.7.5. Evanston, Illinois: Northwestern University. Retrieved from http://cran.r-project.org/package=psych

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, *86*(2), 420–428. doi:10.1037/0033-2909.86.2.420

Soares, A. P., Machado, J., Costa, A., Iriarte, Á., Simes, A., Almeida, J. J. de, … Perea, M. (2015). On the advantages of word frequency and contextual diversity measures extracted from subtitles: The case of Portuguese. *The Quarterly Journal of Experimental Psychology*, *68*(4), 680–696. doi:10.1080/17470218.2014.964271

Treffers-Daller, J. (2013). Measuring lexical diversity among L2 learners of French: An exploration of the validity of D, MTLD and HD-D as measures of language ability. In S. Jarvis & M. Daller (Eds.), *Vocabulary knowledge: Human ratings and automated measures* (pp. 79–103). Amsterdam: John Benjamins. doi:10.1075/sibil.47

Treffers-Daller, J., Parslow, P., & Williams, S. (2016). Back to basics: How measures of lexical diversity can help discriminate between CEFR levels. *Applied Linguistics*, 1–27. doi:10.1093/applin/amw009

Tweedie, F. J., & Baayen, R. H. (1998). How variable may a constant be? Measures of lexical richness in perspective. *Computers and the Humanities*, *32*(5), 323–352. doi:10.1023/A:1001749303137

van Heuven, W. J. B., Mandera, P., Keuleers, E., & Brysbaert, M. (2014). SUBTLEX-UK: A new and improved frequency database for British English. *Quarterly Journal of Experimental Psychology*, *67*(6), 1176–1190. doi:10.1080/17470218.2013.850521

Vanhove, J. (2018). Using text-based indices to predicting human ratings of the lexical richness of short French, German, and Portuguese texts written by children (technical report). Available from https://osf.io/vw4pc/

Wood, S. (2017). *mgcv: Mixed GAM computation vehicle with automatic smoothness estimation*. R package, version 1.8-22. Retrieved from http://cran.r-project.org/package=mgcv

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives in Psychological Science*, *12*(6), 1100–1122. doi:10.1177/1745691617693393

Yeo, I., & Johnson, R. (2000). A new family of power transformations to improve normality or symmetry. *Biometrika*, *87*, 954–959. doi:10.1093/biomet/87.4.954

Yule, G. U. (1944). *The statistical study of literary vocabulary*. Cambridge: Cambridge University Press.